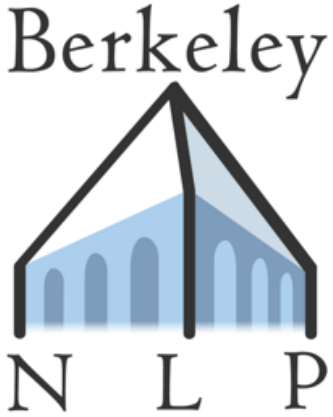
Describing Differences between Text Distributions with Natural Language



Ruiqi Zhong, Charlie Snell, Dan Klein, Jacob Steinhardt

Difference between Distributions

D_1

- Ma mère m’a emmené à l’hôpital.
- J'ai 10 $. Je dépense 3 $ sur un livre.
- Le gouvernement n'a pas réussi à localiser les suspects.

D_2

- My mom and I were best friends.
- Lucy and Peter co-authored a paper.
- I called her to explain why I did badly on the test.

Difference between Distributions

D_1

- Ma mère m’a emmené à l’hôpital.
- J'ai 10 $. Je dépense 3 $ sur un livre.
- Le gouvernement n'a pas réussi à localiser les suspects.

D_2

- My mom and I were best friends.
- Lucy and Peter co-authored a paper.
- I called her to explain why I did badly on the test.

s = “D_1 contains more French sentences compared to D_2”

Tell the Difference!

D_1

- Pieck rescued Gabi from the dungeon and transformed into a Titan afterwards.
- All four of my maternal and fraternal grandparents are professors, and that's why I'm determined to become a prof as well.
- My mom took me to the hospital, and the nurse said that she has never seen this symptom before.
- I was really fortunate to be advised Prof. McKeown and Prof. Hirschberg at Columbia on NLP research, and Prof. Andoni on Theoretical computer science.
- Historia was born as the illegitimate and unrecognized daughter of Rod Reiss. Her mother, Alma, was a servant in his household.
- I called her to explain what happened to her aunt.
- It's quite ironical that such a centralized government fail to locate the suspects who gravely injured those girls earlier this month.

D_2

- She carried a total of eight torpedoes. Her deck was reinforced to enable her to lay a minefield.
- My mom and I were best friends and we used to hunt together.
- Lucy and Peter co-authored a paper on machine learning but got a really bad review.
- I called her to explain why I did really badly on the test.
- Adding to Historia's isolation, the other children outside the estate would throw rocks at her, and she was not allowed to leave.
- Bentham defined as the "fundamental axiom" of his philosophy the principle that "it is the greatest happiness of the greatest number that is the measure of right and wrong."
- Large language models advanced the state of the art by quite a lot but there are still rooms for improvements.
- After 10 years of lockdown due to the pandemics, I finally saw my grandfather — I thought I might never see him again.

Example Insights

Example Insights

- The test distribution involves more formal writing than the training distribution.
  D_1 s D_2

Example Insights

- The test distribution involves more formal writing than the training distribution.
  D_1 s D_2
- A text cluster contains more sports-related articles than other clusters.
  D_1 s D_2

Example Insights

- The test distribution involves more formal writing than the training distribution.
  D_1 s D_2
- A text cluster contains more sports-related articles than other clusters.
  D_1 s D_2
- Public opinions from this year are more optimistic about the pandemic than last year.
  D_1 s D_2

What is a Good Description?

A good description helps humans tell D_1 and D_2 and apart.

Verifying a good description

s = “Samples from D_1 are more positive than those from D_2”

Verifying a good description

s = "Samples from D_1 are more positive than those from D_2"

$x_a \sim D_a$ "This paper proposes an impactful task …"

$x_b \sim D_b$ "The approach of this paper is too trivial."

Verifying a good description

s = "Samples from D_1 are more positive than those from D_2"

$x_a \sim D_a$

"This paper proposes an impactful task …"

$x_b \sim D_b$

"The approach of this paper is too trivial."

Human Classifies

Verifying a good description

s = "Samples from D_1 are more positive than those from D_2"

$x_a \sim D_a$

$x_b \sim D_b$

"This paper proposes an impactful task …"

"The approach of this paper is too trivial."

Human Classifies

I think x_a is from D_1 and x_b is from D_2

Verifying a good description

s = "Samples from D_1 are more positive than those from D_2"

$x_a \sim D_a$

$x_b \sim D_b$

"This paper proposes an impactful task …"

"The approach of this paper is too trivial."

Human Classifies

I think x_a is from D_1 and x_b is from D_2

Loss(*s*): Repeat 100 times and calculate human classification error rate.

Verifying a good description

s = "Samples from D_1 are more positive than those from D_2"

$x_a \sim D_a$

$x_b \sim D_b$

"This paper proposes an impactful task …"

"The approach of this paper is too trivial."

Human Classifies

I think x_a is from D_1 and x_b is from D_2

Loss(*s*): Repeat 100 times and calculate human classification error rate.

~$10 each single description.

Searching for the Best Description

Searching for the Best Description

- Search for the best description that helps humans tell D_1 and D_2 apart.

Searching for the Best Description

- Search for the best description that helps humans tell D_1 and D_2 apart.
- Naive implementation:

Searching for the Best Description

- Search for the best description that helps humans tell D_1 and D_2 apart.
- Naive implementation:
  - Enumerate all natural language strings.

Searching for the Best Description

- Search for the best description that helps humans tell D_1 and D_2 apart.
- Naive implementation:
  - Enumerate all natural language strings.
  - For each string, verify its quality by asking humans to use it to classify on 100 sample pairs.

Searching for the Best Description

- Search for the best description that helps humans tell D_1 and D_2 apart.
- ~~Naive~~ Practical implementation:
  - ~~Enumerate all natural language strings.~~

    Fine-tune GPT-3 to propose promising candidate descriptions.
  - ~~For each string, verify its quality by asking humans to use it to classify on 100 sample pairs.~~

    Fine-tune model to simulate human classification.

Exposing Dataset Flaws

Exposing Dataset Flaws

- Machine learning models might pick up shallow undesirable correlations.

Exposing Dataset Flaws

- Machine learning models might pick up shallow undesirable correlations.
- Binary classification: Spam vs. Non-Spam

D_1 D_2

Exposing Dataset Flaws

- Machine learning models might pick up shallow undesirable correlations.
- Binary classification: Spam vs. Non-Spam
  D_1 D_2
- "D_1 contains more spam" / "D_1 contains more hyperlinks"

Exposing Dataset Flaws

- Machine learning models might pick up shallow undesirable correlations.
- Binary classification: Spam vs. Non-Spam

  D_1 D_2
- "D_1 contains more spam" / "D_1 contains more hyperlinks"
- RoBERTa fine-tuned on this dataset classifies a message as spam whenever it sees a hyperlink!!!

In Our Paper

In Our Paper

- A benchmark with 54 real-world distribution pairs with known differences.

In Our Paper

- A benchmark with 54 real-world distribution pairs with known differences.
- An automatic data generation method to fine-tune GPT-3.

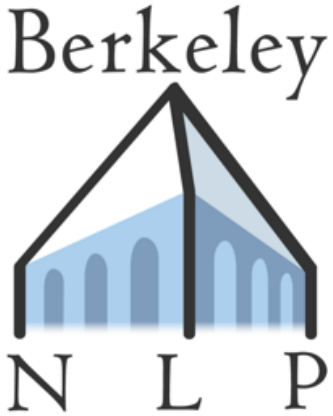
In Our Paper

- A benchmark with 54 real-world distribution pairs with known differences.
- An automatic data generation method to fine-tune GPT-3.
  - GPT-3 0-shot: 7%.
  - Fine-tuned GPT-3 with re-ranking: 61%.
- Applications:
  - Summarize unknown tasks.
  - Describe distribution shifts.
  - Expose dataset flaws.
  - Label text clusters.

In Our Paper

- A benchmark with 54 real-world distribution pairs with known differences.
- An automatic data generation method to fine-tune GPT-3.
  - GPT-3 0-shot: 7%.
  - Fine-tuned GPT-3 with re-ranking: 61%.
- Applications:
  - Summarize unknown tasks.
  - Describe distribution shifts.
  - Expose dataset flaws.
  - Label text clusters.

Our system finds dataset properties we were unaware of before!!



Thanks!